IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13426

ScoreLens: Automated Student Result Processing and Analysis from Academic Gazette PDF

Kanda Kumaran Thevar[1],

Samiksha Pawar [2], Trisha Pashte[3] , Kaveri Shivkar[4], Priyanka Khot[5]

Lecturer, Department of Information Technology, BVIT, Navi Mumbai, India[1]

Student, Department of Information Technology, BVIT, Navi Mumbai, India [2]

Student, Department of Information Technology, BVIT, Navi Mumbai, India [3]

Student, Department of Information Technology, BVIT, Navi Mumbai, India [4]

Student, Department of Information Technology, BVIT, Navi Mumbai, India [5]

Abstract: The management and analysis of student academic results in diploma education systems often rely on manually processing gazette PDF documents, which is time-consuming and susceptible to human error. This paper presents the design and implementation of an automated system for extracting, structuring, and analyzing student results specifically based on MSBTE (Maharashtra State Board of Technical Education) standard formats and academic requirements. The proposed system efficiently processes structured PDF gazette files, extracts relevant student information, and converts it into a well-organized format suitable for analysis.

The system utilizes PDF parsing techniques through PyMuPDF to accurately extract data from MSBTE-formatted documents, followed by data structuring and processing using Python libraries such as Pandas and OpenPyXL. It performs comprehensive result analysis, including subject-wise performance evaluation, overall result computation, and structured report generation in Excel format. By adhering to MSBTE standards, the system ensures compatibility, consistency, and reliability in handling academic records.

Unlike existing approaches that rely heavily on OCR and machine learning techniques for unstructured data processing , the proposed system focuses on a domain-specific, rule-based approach optimized for structured academic documents. Experimental results demonstrate improved efficiency, reduced manual effort, and high accuracy in data extraction and analysis. The system provides a scalable solution for academic institutions and can be extended in the future to support unstructured formats using advanced techniques such as OCR and natural language processing.

Keywords: Student Performance Analysis, Document Parsing, PDF Data Extraction, Educational Data Mining, Automated Academic Systems, Data Structuring, Result Analytics, Python-Based System, MSBTE Gazette Processing, Information Extraction.

INTRODUCTION

The rapid growth of digital technologies has significantly transformed the way academic data is generated and managed in educational institutions. Student results are commonly published in the form of gazette PDF documents, which contain large volumes of structured information. Although these documents are digitally available, the process of extracting, organizing, and analyzing the data remains largely manual in many institutions. This manual approach is time-consuming, inefficient, and prone to human errors, especially when dealing with a large number of students.

In academic environments such as the Maharashtra State Board of Technical Education (MSBTE), result gazettes follow a standardized format designed primarily for human readability rather than automated processing. As a result, educators and administrators are required to manually extract student details, calculate totals and percentages, and prepare reports for analysis. This not only increases workload but also creates challenges in maintaining accuracy and consistency in result processing.

To overcome these limitations, there is a need for an automated system that can efficiently process and analyze academic result data. This paper presents ScoreLens, an automated result analyzer system designed to extract, structure, and analyze student results from gazette PDF documents. The system utilizes document parsing techniques to extract

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13426

relevant information such as student details, subject marks, and result status, and converts it into a structured format suitable for further processing.

The extracted data is then analyzed to generate meaningful insights into student performance, including subject-wise analysis, overall result evaluation, and identification of performance trends. The system also generates structured Excel reports, enabling easy interpretation and accessibility of data. By automating the entire workflow, the system reduces manual effort, improves accuracy, and enhances efficiency in result analysis.

The proposed system is specifically designed to work with MSBTE standard formats, ensuring reliable and accurate data extraction. It is implemented as a standalone Python-based application, making it easy to use and deploy without requiring complex infrastructure. Additionally, the system is scalable and can be extended in the future to support multiple academic formats, web-based interfaces, and advanced analytical features.

In summary, the ScoreLens Result Analyzer provides an effective solution for automating the processing and analysis of academic results, thereby improving data management and supporting better decision-making in educational institutions.

LITERATURE SURVEY

In today's data-driven educational environment, institutions generate large volumes of academic data, including student marks, attendance records, and performance reports. Traditional result analysis methods primarily rely on manual data entry and spreadsheet-based processing, which are time-consuming, error-prone, and inefficient when handling large datasets. With advancements in technologies such as Optical Character Recognition (OCR), Natural Language Processing (NLP), Machine Learning (ML), and Data Analytics, there has been a significant shift toward automation in educational systems. These technologies enable automatic data extraction, intelligent evaluation, and effective visualization of academic results, thereby improving overall efficiency and accuracy.

A notable contribution in this domain is the work by Dr. Anindita Khade et al., which introduces the QualiScore system for automated evaluation of subjective answers using NLP and machine learning [1]. The system utilizes semantic similarity techniques and models such as Gradient Boosting for score prediction, along with OCR integration for processing handwritten responses. While this approach enhances answer evaluation, it does not provide a complete solution for result analysis and visualization.

OCR technology plays a crucial role in converting scanned documents into machine-readable text by recognizing pixel patterns, thereby enabling digitization and data extraction [2]. Practical implementations using tools such as Tesseract OCR have demonstrated the ability to process answer sheets and automate mark calculation [3]. However, these systems often face challenges in terms of accuracy, particularly when dealing with handwritten data and complex document layouts.

In addition to text extraction, significant research has been conducted in the area of structured data extraction from documents, particularly tables. Techniques such as detection, segmentation, and interpretation are widely used to extract tabular data from PDF files. End-to-end table extraction approaches improve accuracy by analyzing both the structure and relationships within tables [4]. Methods that utilize visual separators, such as whitespace and ruling lines, have proven effective in identifying tables in complex PDF layouts [5]. Furthermore, heuristic approaches combined with OCR techniques have been used to reconstruct tabular structures efficiently [6]. Machine learning-based methods further enhance this process by identifying table elements and understanding document context, although they require higher computational resources.

Modern systems have also focused on integrating data processing and visualization into unified platforms. Python-based frameworks and tools such as Streamlit enable the development of complete pipelines for data extraction, processing, and visualization, allowing the generation of dashboards and analytical insights [8], [9]. These systems demonstrate that automation not only improves efficiency and accuracy but also enhances decision-making through effective data visualization.

Despite these advancements, several limitations still exist in current solutions. Many systems lack full integration of OCR, NLP, ML, and data analytics into a single platform. There is also limited focus on end-to-end automation, with most approaches addressing only specific stages of the workflow. Additionally, existing systems often do not support MSBTE-based formats for result processing and export, and many solutions are domain-specific with limited scalability.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13426

To address these gaps, the proposed system, ScoreLens, is designed as an integrated platform that combines multiple technologies into a unified solution. The system utilizes OCR tools such as Tesseract and PyMuPDF for extracting data from PDF documents, along with data processing libraries like Pandas and OpenPyXL for efficient data handling. It provides structured and exportable reports in Excel format, ensuring usability and accessibility.

Unlike existing approaches, ScoreLens focuses on end-to-end automation, a user-friendly interface, accurate and fast processing, and a scalable architecture. By integrating extraction, processing, analysis, and reporting into a single system, it aims to provide a comprehensive solution for automated result analysis.

Overall, the literature indicates a clear transition from manual result processing methods to intelligent automated systems. While technologies such as OCR, NLP, machine learning, and data analytics have significantly improved the efficiency of educational systems, existing solutions still lack full integration and usability. The proposed system aims to bridge this gap by delivering a comprehensive, automated, and efficient platform for academic result analysis..

PROBLEM STATEMENT

Educational institutions commonly rely on manual or semi-automated methods to process and analyze student results from gazette PDF documents. These approaches are time-consuming, prone to human errors, and inefficient for handling large datasets. Additionally, existing systems lack proper data structuring, automated analysis, and support for MSBTE-specific formats, making it difficult to generate accurate insights and manage academic records effectively. Therefore, there is a need for an automated system that can efficiently extract, process, and analyze student result data with high accuracy and minimal manual intervention.

SCOPE OF PROJECT

The scope of the proposed system includes the development of an automated solution for extracting, processing, and analyzing student results from MSBTE gazette PDF documents. The system focuses on structured data extraction, result computation, performance analysis, and generation of organized Excel reports. It also includes basic data visualization to support interpretation of results. The implementation is limited to a standalone Python-based application designed for MSBTE formats, with potential for future enhancements such as multi-format support, web-based deployment, and advanced analytics.

PROPOSED METHODOLOGY

The proposed methodology of the ScoreLens Result Analyzer focuses on developing an automated system for processing gazette PDF files and converting them into structured and analyzable Excel reports. The system begins by accepting a gazette PDF as input, from which relevant student data such as name, enrollment number, marks, and result status is extracted using PDF parsing techniques.

The extracted data undergoes preprocessing to remove inconsistencies and is then structured into organized formats based on categories such as department, semester, and scheme. Subsequently, the system performs analysis to generate insights including topper lists, failure analysis, and overall performance summaries.

Finally, the processed data is exported into well-structured Excel files containing multiple categorized sheets for easy interpretation. The system follows a modular architecture comprising input, processing, analysis, and output components, ensuring efficient data flow, accuracy, and scalability.

SYSTEM DESIGN

System design plays a crucial role in defining the structure, functionality, and data flow of the ScoreLens Result Analyzer. It acts as a blueprint that translates system requirements into a structured architecture, ensuring efficient, reliable, and maintainable implementation. The system is designed using a modular approach to simplify development and enhance flexibility, where each module performs a specific function.

The overall architecture of the system is divided into three primary components: Input, Processing, and Output. The input module is responsible for accepting gazette PDF files and validating the data. The processing module handles data extraction, cleaning, and computation, including calculation of total marks, percentages, grades, and performance

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13426

analysis. The output module presents the results in a structured format, including tabular data and graphical representations, ensuring clarity and usability.

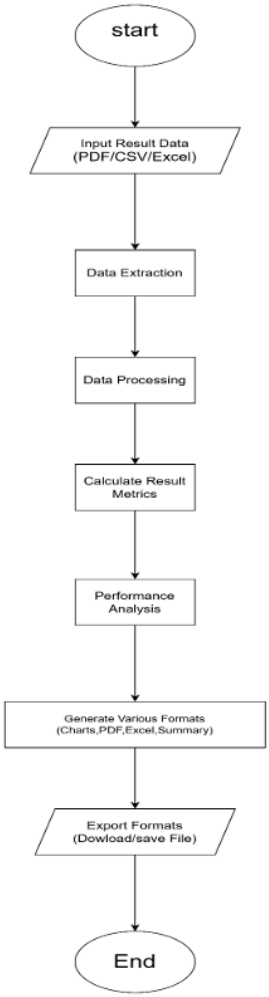


Fig. 1 Flow Diagram

The system follows a simple data flow:
Input → Processing → Output, which ensures smooth and efficient handling of data throughout the pipeline.

To represent system functionality, a flow-based approach is used where the process begins with data input, followed by validation and processing, including calculations and analysis, and ends with result generation and display. This structured workflow helps in reducing errors and improving system performance.

The Data Flow Diagram (DFD) illustrates the movement of data between the user and the system. At the context level, the user provides input to the system, which processes the data and generates output. At a detailed level, the process involves data input, processing, result generation, and final output presentation.

The system uses a basic structured data storage approach, where student information such as name, roll number, subject marks, total marks, percentage, and grade is organized in tabular format. This ensures easy data processing, retrieval, and consistency.

The user interface is designed to be simple and user-friendly, allowing users to upload files, process data, view results, and analyze outputs without requiring technical expertise. This improves usability and enhances the overall user experience.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

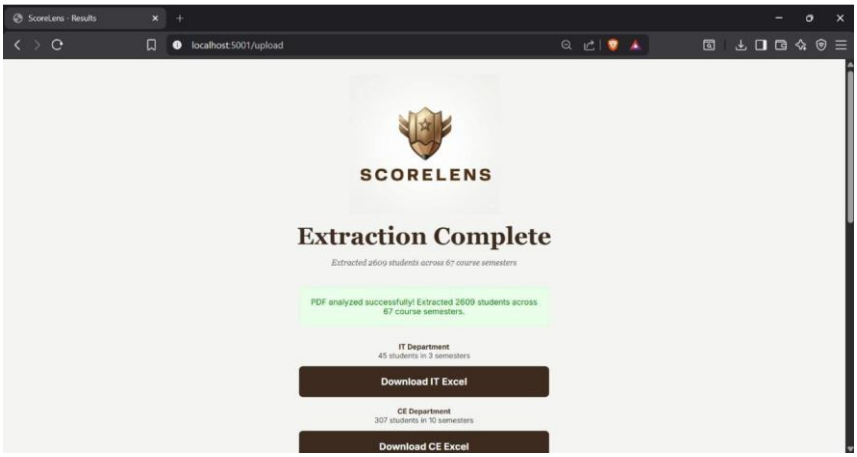
DOI: 10.17148/IARJSET.2026.13426

Overall, the modular and structured design of the system ensures efficient data handling, reduced complexity, and ease of future enhancements.

RESULTS AND DISCUSSION

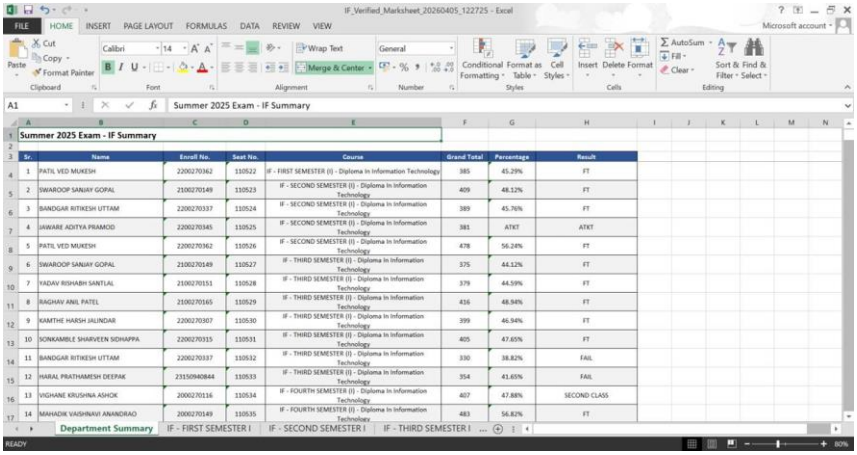
The ScoreLens Result Analyzer system was implemented and tested using multiple MSBTE gazette PDF files to evaluate its performance, accuracy, and efficiency. The system successfully processed the input PDF documents, extracted relevant student data, and generated structured Excel reports without manual intervention.

Fig. 2 Prompt after Extraction is completed



The system demonstrated accurate extraction of key student information, including name, enrollment number, subject-wise marks, percentage, and result status. The extracted data was effectively cleaned and organized into structured formats such as department-wise, semester-wise, and scheme-wise classifications. The generated Excel reports contained multiple sheets, including detailed student records, topper lists, and failure analysis, which improved data readability and usability.

Fig. 3 Exported Excel Sheet



The processing time was significantly reduced compared to manual methods, as the system automated the entire workflow from data extraction to report generation. Even for large gazette files, the system maintained consistent performance and produced results in a short time. The modular design ensured smooth data flow and minimized processing errors.

The results indicate that the proposed system effectively addresses the limitations of traditional manual result processing methods. By automating data extraction and analysis, the system reduces human effort and eliminates common errors associated with manual calculations and data entry. The structured output in Excel format enhances accessibility and allows easy further analysis.

One of the key strengths of the system is its ability to handle MSBTE-standard gazette formats with high accuracy due to its rule-based extraction approach. The categorization of data into multiple analytical views, such as toppers and failure lists, provides meaningful insights that support academic decision-making.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13426

However, the system has certain limitations. It is currently dependent on the predefined format of MSBTE gazette PDFs, which may restrict its applicability to other formats without modification. Additionally, while the system performs structured data extraction efficiently, it does not yet support advanced analytics or unstructured data processing.

Overall, the ScoreLens Result Analyzer demonstrates significant improvements in efficiency, accuracy, and data management compared to traditional methods. The system provides a practical and scalable solution for academic institutions and lays the foundation for future enhancements, including support for multiple formats and integration of advanced analytical techniques.

APPLICATIONS

The ScoreLens Result Analyzer system has several practical applications in academic environments where efficient handling of student result data is required. It can be used by educational institutions such as diploma colleges and technical boards to automate the processing and analysis of gazette PDF results.

The system is useful for teachers and administrators to quickly generate structured reports, identify toppers and failures, and analyze subject-wise and overall student performance. It also supports decision-making by providing clear insights into academic trends and student outcomes.

Additionally, the system can be applied in examination departments for efficient result management and record keeping. It reduces manual workload, improves accuracy, and ensures organized data storage. With further enhancements, it can be extended to support multiple universities, integrated academic management systems, and web-based result analysis platforms.

CONCLUSION

This paper presented the design and implementation of the ScoreLens Result Analyzer, an automated system developed to efficiently process and analyze student results from MSBTE gazette PDF documents. The system successfully addresses the limitations of traditional manual methods by automating data extraction, processing, and report generation, thereby reducing time, effort, and the possibility of human errors.

The proposed system demonstrates effective performance in extracting structured student data, organizing it into meaningful formats, and generating comprehensive Excel reports for analysis. The modular architecture ensures smooth data flow and scalability, while the use of Python-based tools enables efficient handling of large datasets. The system also enhances usability by providing clear insights such as subject-wise performance, topper lists, and overall result trends, which support better academic decision-making.

Although the current implementation is limited to MSBTE-specific formats, it provides a strong foundation for future enhancements. The system can be extended to support multiple university formats, web-based deployment, and advanced analytical techniques such as machine learning and data visualization.

In conclusion, the ScoreLens Result Analyzer offers a practical, accurate, and scalable solution for automating academic result analysis, contributing to improved efficiency and data management in educational institutions.

ACKNOWLEDGMENT

The authors express their sincere gratitude to their Guide, Mr. Kanda Kumaran Thevar for his valuable guidance, Project Co-Ordinator Mr. Sandeep Shinde for his valuable support throughout the project. We also thank the Head of Department, Mr. Ranjeet Pawar, and the Principal, Mr. P. N. Tandon, for their encouragement and support.

We extend our appreciation to Dr. Anindita Achint Khade, Mr. Madhavan Naikar, and Mr. Gaurav Jadhav for their assistance during the literature survey. Finally, we thank the Department of Information Technology and all teaching and non-teaching staff for their support in successfully completing this project.

REFERENCES

[1]. G. Jadhav et al., “Design of an Auto Evaluation Model for Subjective Answers using NLP and Machine Learning Techniques – QualiScore,” under guidance of Dr. Anindita Khade, 2023–24.

[2]. Mukherjee, Sumita, Hritik Tyagi, Purushautam Tyagi, Nikita Singh, and Shraddha Bhardwaj. "OCR using python and its application." Journal of Advanced Zoology 44 (2023): 1083.

DOI: 10.17148/IARJSET.2026.13426

[3]. Saravanan, Kalaimathi, Chang Choon Chew, Kim Gaik Tay, Sie Long Kek, and Audrey Huong. "Exam Marks Summation App Using Tesseract OCR in Python." International Journal of Integrated Engineering 14, no. 3 (2022): 102-110.

[4]. e Silva, Ana Costa, Alípio M. Jorge, and Luís Torgo. "Design of an end-to-end method to extract information from tables." International Journal of Document Analysis and Recognition (IJDAR) 8, no. 2 (2006): 144-171.

[5]. Fang, Jing, Liangcai Gao, Kun Bai, Ruiheng Qiu, Xin Tao, and Zhi Tang. "A table detection method for multipage pdf documents via visual seperators and tabular structures." In 2011 International Conference on Document Analysis and Recognition, pp. 779-783. IEEE, 2011.

[6]. Vasileiadis, Manolis, Nikolaos Kaklanis, Konstantinos Votis, and Dimitrios Tzovaras. "Extraction of tabular data from document images." In Proceedings of the 14th International Web for All Conference, pp. 1-2. 2017.

[7]. Sarangpure, Nikhilesh, Vipul Dhamde, Ankita Roge, Janhawi Doye, Shivam Patle, and Sukhad Tamboli. "Automating the machine learning process using PyCaret and Streamlit." In 2023 2nd International Conference for Innovation in Technology (INOCON), pp. 1-5. IEEE, 2023.

[8]. A Python Framework For Academic Data Analysis And Visualization". 2025. International Journal of Engineering Research and Science & Technology 21 (4): 83-88.

